DOCUMENT RESUME

ED 315 448 TM 014 475

AUTHOR Perry, Lucille N.
TITLE The Use of Structure Coefficients in Regression Research.
PUB DATE Jan 90
NOTE 16p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, January 25-27, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Correlation; Educational Research; Heuristics; Multivariate Analysis; Predictive Measurement; *Regression (Statistics); *Research Methodology; Test Interpretation
IDENTIFIERS *Structure Coefficients

ABSTRACT

It is recognized that parametric methods (e.g., t-tests, discriminant analysis, and methods based on analysis of variance) are special cases of canonical correlation analysis. In canonical correlation it has been argued that structure coefficients must be computed to correctly interpret results. It follows that structure coefficients may be useful in interpreting regression results. A small data set for a random sample of 12 of 367 participants in a study of cardiovascular risk factors is used to demonstrate structure coefficients and their use. This example and one drawn from educational research on testwiseness illustrate that failure to utilize structure coefficients can lead to serious distortions. The relationship between beta weights and structure coefficients is examined. The comparison of beta weights and structure coefficients provides important information about the predictive model. Both are important to the researcher interpreting regression results. Five data tables illustrate the examples. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED315448

strcoeff.wp 1/16/90

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Lucille N. Perry
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Use of Structure Coefficients in Regression Research

Lucille N. Perry

University of New Orleans 70148

Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX, January 27, 1990.

TM014475

BEST COPY AVAILABLE

ABSTRACT

It is well known that all parametric methods, e.g., t-tests, ANOVA, ANCOVA, MANOVA, MANCOVA, and discriminant analysis, are special cases of canonical correlation analysis (Knapp, 1978; Thompson, 1984, 1988). In the canonical case it has been emphatically argued by various researchers that structure coefficients must be computed in order to correctly interpret results (Kerlinger & Pedhazur, 1973, p. 344; Levine, 1977, p. 20; Meredith, 1964, p. 55), although Harris (1989) disagrees.

Since regression is a special case of canonical analysis (Thompson & Borrello, 1985), it stands to reason that structure coefficients may also be very useful in interpreting regression results. Indeed, Cooley and Lohnes (1971, p. 55) make just such an argument.

The purpose of this paper is to explain what structure coefficients are through the use of a small data set to make the discussion concrete. It is shown that the failure to utilize structure coefficients during interpretations in educational research can lead to serious distortions. Ways of computing structure coefficients using commonly available statistical packages are presented.

Prediction or predictive power is one of the concepts most emphasized in the study of statistics. Scientists and researchers are intent on learning about the relationships of the elements in the surrounding physical or social reality. Researchers devise methods to measure and record in consistent standardized increments those traits of primary interest. These measurements or scores called data are then mathematically arranged into some sort of model explaining how the value for a certain variable, a dependent variable, changes in relation to other events occurring in the same environment. In this way scientists, educators, researchers, and statisticians, given a set of data, can predict what would happen in the future assuming the occurrence of certain combinations of events, at least with a certain degree of accuracy. In effect, the laws of statistics and mathematics are the tools of this prediction "magic", and as any practical person might argue, the better equipped and skilled we are at handling our tools, the more informed and accurate will be our predictions.

Commonly used statistical tools are analytic methods such as multiple regression, t tests, ANOVA, ANCOVA, MANOVA, MANCOVA and discriminant analysis. However, all parametric methods are all special cases of a more general analytic method called canonical analysis (Knapp, 1978; Thompson, 1988). All these methods are used to build mathematical models of variable relatedness.

All parametric methods also yield some estimate of the predictive power of a given model. For example, in multiple correlation analysis this "effect size" estimate ranges between

zero and 100%, and is the squared multiple correlation coefficient, R^2. In ANOVA, an effect size called the correlation ratio or eta^2 is computed by dividing the sum of squares explained for a given effect by the sum of squares total.

In the real world to which the researcher wishes to generalize, it is often untenable to assume that the predictor variables that have been chosen are mutually exclusive or uncorrelated with each other. Statisticians call this situation "collinearity." In fact, some investigators purposely choose predictor measures that are intuitively or empirically expected to be collinear. For example, researchers often elect to "multioperationalize" predictor variables that they believe may be particularly important or which may be particularly difficult to measure in a reliable and valid fashion.

The use of a mathematical model is most explicit in multiple regression analysis, although the same basic model is implicit in all parametric analytic methods. The researcher wants to optimally weight observed predictor variable scores (e.g., X_1 and X_2) so that the predicted score, Y^*, for each person comes as close as possible to each person's actual dependent variable score, Y. Thus, the model takes the form:

$$Y \text{ <---- } Y^* = a + b_1(X_1) + b_2(X_2).$$

The weight a equals the Y^* score when X_1 and X_2 both equal zero. The weights (a, b_1, and b_2) are "optimal" when <u>any other</u> weights make the estimated scores (Y^*) and the actual dependent variable scores (Y) further apart. This distance between each person's actual (Y)

score and predicted score (Y^*) is each person's error or e score.

If each person's predicted score exactly equals each person's actual Y score, each person's e score will be zero. If the predictor variables are of no use at all, b_1 and b_2 will both equal zero, a will equal the mean of Y, and each person's predicted score (Y^*) will equal the mean of Y also. The bivariate correlation between the Y and Y^* scores is called R, the multiple correlation coefficient. If all the e scores equal 0, R will be its maximum possible value, 1.0. If all the e scores equal the mean of Y, R will be its minimum possible value, 0.0.

Interpreting Models When R<>0

When the multiple correlation coefficient does not equal R, the researcher usually wants to know which predictor variable or combination of predictors contributes to the predictive power of the model. One obvious place to look is at the optimal multiplicative weights, the b weights, b_1 and b_2. If one of them is zero, the effect of the weighting is to totally remove that predictor from influencing the estimation of Y^*, thus such predictors have no unique predictive power.

When none of the b weights are zero, one might think that the magnitude of the b weights could be compared to determine which predictors are most useful. The problem with this logic is that the b weights are sensitive to three influences: (a) the correlation between each predictor and Y, (b) the correlation among the predictors and each other, and (c) the relative "spreadoutness" or

variability of each predictor in relation to the dependent variable, Y.

With respect to this last influence, since the focus of the model is to predict Y, one purpose of the b weights is to convert the "spreadoutness" of the predictor into the same metric as the "spreadoutness" of the dependent variable. For example, even if a predictor such as GRE score was highly related to a dependent variable such as GPA, GRE (ranging between 400 and 1600) would still have a very small b weight (probably a near zero value) so as to convert the GRE scores into the same metric as GPA (ranging between 0 and 4).

Some researchers avoid the influence of this third dynamic, i.e., the influence of the "spreadoutness" of the variables, by converting all the b weights into beta weights, i.e., the weights that would be used if all the variables were in standardized Z score form. The beta weight for a given predictor equals the b weight for that predictor times the ratio of the SD of the predictor to the SD_Y, i.e.,

$$beta_X = b_X \ (SD_X \ / \ SD_Y).$$

One might think that this takes care of any problems with a confounded interpretation of the weights, and that the importance of the predictors could be determined simply by consulting the beta weight values. Unfortunately, although the influence of (c) variable "spreadoutness" has been removed from the beta weights, two influences--(a) correlation of each predictor with Y and (b) correlation of the predictors with each other--still affect the

beta weights. Thus, the researcher still is not sure what exactly is contributing to the large or small beta weights for given predictors.

What is needed is a measure of variable importance that is only sensitive to the predictive ability of a given variable. Since R equals the bivariate correlation between Y and Y^*, logically, what is needed is something that will inform the researcher regarding the nature and substance of Y^*. Logically, this can be evaluated by correlating each predictor variable with the Y^* scores. Such correlations are called **structure coefficients.**

Heuristic Example #1

A small heuristic example might be useful in illustrating the importance of structure coefficients in regression research. Table 1 presents a data set derived from the "Heart Smart" study of cardiovascular risk factors in children (Downey et al., 1987, 1988). HDL cholesterol, the so-called "good" cholesterol, is the dependent variable (Y) in this example. The predictors are: (a) time in seconds to run or walk one mile (MILESEC); (b) and average of six systolic blood pressure measurements (SYSTOLAV); and (c) ponderosity, the ratio of weight in kilograms to the cubed value of height in centimeters (POND).

INSERT TABLE 1 ABOUT HERE.

The R for these data is 0.58405. The beta weights and structure coefficients from the analysis are presented in Table 2.

It is noteworthy that the person interpreting only the beta weights would conclude that POND (beta = 0.0096) has almost no predictive power. Yet, as the structure coefficients in Table 2 indicate, ponderosity (|structure r| = |-.278|) has roughly the same ability to delineate Y^* as does MILESEC (structure r = +.2395).

INSERT TABLE 2 ABOUT HERE.

Table 3 illustrates the calculation of various results associated with the analysis. It is noteworthy that the table implies another formula for estimating the structure coefficients. The structure coefficient for a given predictor always equals the bivariate correlation between the predictor and Y divided by the multiple correlation coefficient, R. For example, the structure coefficient for MILESEC (+.2395) equals +.1399/+.5840.

INSERT TABLE 3 ABOUT HERE.

Heuristic Example #2

Table 4 presents the correlation matrix associated with a second example reported by Thompson and Borrello (1985). The dependent variable in this example involves scores on a measure of testwiseness (Tst-W). There are eight predictor variables. As suggested by the results reported in Table 5, the person interpreting only beta weights would erroneously conclude that Intuition (beta weight = +.05933) has virtually no predictive power. But the structure coefficient for Intuition (+.78320)

suggests that this predictor, taken alone, is the second best predictor in the whole set of eight predictors!

INSERT TABLES 4 AND 5 ABOUT HERE.

The example illustrates that the comparison of beta weights and structure coefficients provides important information about the predictive model. When the predictor variables are perfectly uncorrelated, each predictor's beta weight will equal its correlation with Y, and the predictors will be ordered exactly the same by both beta weights and structure coefficients with respect to predictive power. When variables are ordered differently by the two sets of coefficients, as in the present example, the results indicate that collinearity among the predictors may influence result interpretation.

Summary

Regression analysis is frequently used in educational research. But often researchers interpret their results solely by consulting R and the beta weights for the predictor variables. The present examples make clear that such practice can lead to seriously erroneous interpretations. Since the weights are used to define the predicted dependent variables scores, Y^*, it does make sense to consult the weighting coefficients, and the beta weights are appropriate for this purpose since the influence of variability of the variables is removed from the beta weights, as noted previously.

However, it is important to also consult either (a) the structure coefficients for the predictors or (b) the bivariate correlation coefficient involving each predictor and the dependent variable, Y. These two sets of results will always order the predictors identically, so either of these two sets of results can be interpreted. But it is important that one or the other of these two sets of results be interpreted, and that interpretation not be based solely upon the beta weights.

Attempting to interpret the "magic" of prediction without checking the "test light" of the structure coefficients along with beta weights is tantamount to choosing astigmatism over corrected sight. Both beta weights and structure coefficients are important to the educational researcher who wishes to correctly interpret regression results.

References

Cooley, W.W., & Lohnes, P.R. (1971). Multivariate data analysis. New York: Irvington.

Downey, A.M., Frank, G.C., Weber, L.S., Harsha, D.W., Virgilio, S., Franklin, F.A., & Berenson, G.S. (1987). Implementation of "Heart Smart": A cardiovascular school health promotion program. Journal of School Health, 57, 98-104.

Downey, A.M., Virgilio, S., Serpas, D.C., Nicklas, T.A., Arbeit, M.L., & Berenson, G.S. (1988). "Heart Smart"--A staff development model for a school based cardiovascular health intervention. Health Education, 19(5), 64-71.

Harris, R.J. (1989). A canonical cautionary. Multivariate Behavioral Research, 24, 17-39.

Kerlinger, F.N., & Pedhazur, E.J. (1973). Multiple regression in behavioral research. New York: Holt, Rinehart and Winston.

Knapp, T.R. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, 85, 410-416.

Levine, M.S. (1977). Canonical analysis and factor comparison. Newbury Park: SAGE.

Meredith, W. (1964). Canonical correlations with fallible data. Psychometrika, 29, 55-65.

Thompson, B. (1984). Canonical correlation analysis: Uses and interpretation. Newbury Park, CA: SAGE.

Thompson, B. (1988, April). Canonical correlation analysis: An explanation with comments on correct practice. Paper presented

at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 295 957)

Thompson, B., & Borrello, G.M. (1985). The importance of structure coefficients in regression research. Educational and Psychological Measurement, 45, 203-209.

TABLE 1
"Heart Smart"
Random sample of n=12 from N=367

ID	MILESEC	SYSTOLAV	POND	HLDCHOL
1	890	94.0	11.5	80.1
2	1097	108.7	12.0	51.1
3	1300	97.7	13.1	63.3
4	948	90.3	12.6	75.7
5	940	100.7	19.3	61.0
6	760	104.3	14.7	76.0
7	740	95.3	14.2	78.5
8	571	97.7	13.6	54.3
9	748	102.7	10.9	62.2
10	640	96.0	11.4	67.4
11	642	107.0	14.6	34.8
12	957	108.0	15.2	70.8

Note. From Thompson (1990), with permission.

Table 2
Regression beta Weights and Structure Coefficients

	beta Weights	Structure Coefficients
MILESEC	.1688	.2395
SYSTOLAV	-.5700	-.9571
POND	.0096	-.2078

Note. R = .58405

TABLE 3
Correlation Matrix

	SYSTOLAV	POND	HDLCHOL	ERROR	YHAT
MILESEC	.0516	.0463	.1399	.0000[a]	.2395[b]
SYSTOLAV		.2435	-.5590	.0000[a]	-.9571[b]
POND			-.1214	.0000[a]	-.2078[b]
HDLCHOL (Y)				.8117	.5840[c]
ERROR					.0000[d]

Note. From Thompson (1990), with permission.

[a]Since ERROR is that part of Y not explained by or overlapping with the predictors as a set, this part of Y is by definition always uncorrelated with each predictor.

[b]These are the structure coefficients for this example.

[c]The bivariate correlation (r) between YHAT (Y^*) and Y is R.

[d]Since ERROR is the part of Y not explained by the predictors in the form of Y^*, by definition Y^* and ERROR never overlap.

TABLE 4
Correlation Matrix

	Tst-W	Extrv	Sensn	Think	Judgn	Intrv	Intui	Feeln	Perce
Tst-W	1.000								
Extrv	-.051	1.000				-.952			
Sensn	-.368	-.107	1.000				-.855		
Think	-.192	-.090	.284	1.000				-.786	
Judgn	-.117	-.048	.382	.191	1.000				-.970
Intrv	.052	-.952	.083	.061	.024	1.000			
Intui	.313	.114	-.855	-.215	-.390	-.093	1.000		
Feeln	.108	.018	-.205	-.786	-.219	.004	.224	1.000	
Perce	.109	.075	-.384	-.176	-.970	-.040	.419	.225	1.000

Note. From Thompson and Borrello (1985).

TABLE 5
Canonical and Regression Results

Variable	Canonical Function Coefficients (F)	Regression Beta Weights	Structure Coefficients
Extravert	-.37119 x R	-.14824/R = F	.12735
Sensing	-.77559 x R	-.30975/R = F	-.92086
Thinking	-.48355 x R	-.19315/R = F	-.48050
Judging	-.13063 x R	-.05215/R = F	-.29323
Introvert	-.12094 x R	-.04830/R = F	.12926
Intuition	.14855 x R	.05933/R = F	.78320
Feeling	-.27474 x R	-.10972/R = F	.27025
Perceiving	-.21541 x R	-.08603/R = F	.27198

Note. For these data R = Rc = 0.39937. From Thompson and Borrello (1985).